National Aeronautics and Space Administration

HPC and Clouds at the NCCS: Measuring Woody Biomass on South Side of the Sahara at the 40-50 cm scale using AWS

IS&T Colloquium, NASA Goddard Space Flight Center
04 November 2015

Daniel Duffy daniel.q.duffy@nasa.gov and on Twitter @dqduffy
High Performance Computing Lead at the
NASA Center for Climate Simulation (NCCS) – http://www.nccs.nasa.gov and @NASA_NCCS
Goddard Space Flight Center (GSFC) – http://www.nasa.gov/centers/goddard/home/

Rats

Do climate records back that up?

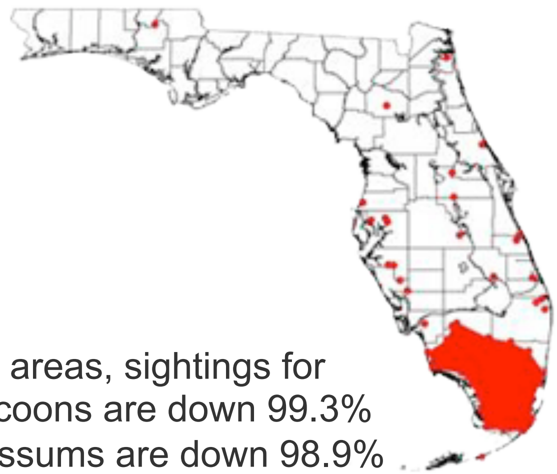
"Analysis of 15 tree-ring records, which document yearly weather conditions, shows that Europe always experienced plague outbreaks after central Asia had a wet spring followed by a warm summer — terrible conditions for black rats, but ideal for Asia's gerbil population. Those sneaky rodents and their bacteria-ridden fleas then hitched a ride to Europe via the Silk Road, arriving on the continent a few years later to wreak epidemiological havoc."

http://www.washingtonpost.com/news/morning-mix/wp/2015/02/24/after-8-centuries-rats-exonerated-in-spread-of-black-death-gerbils-implicated/?tid=hp_mm
http://www.washingtonpost.com/news/morning-mix/wp/2015/06/21/star-teenage-athlete-dies-after-flu-symptoms-turn-out-to-be-plague/

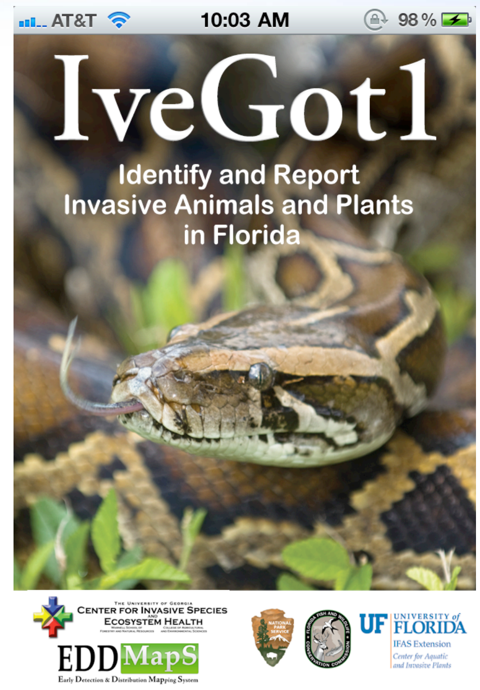
Invasive Species – Pythons in South Florida

Investigating climate suitability and invasive species

Observations of non-native Burmese pythons in Florida – moving Northwest.



In these areas, sightings for
- Raccoons are down 99.3%
- Opossums are down 98.9%
- White-tailed deer are down 94.1%



Source – Proceedings of the National Academy of Sciences, "*Severe mammal declines coincide with proliferation of invasive Burmese pythons in Everglades National Park*"

Estimated Damage from Invasive Species is more than $1.4B

How will changes in climate affect invasive species and how should we prepare for them?
Source:
For more information: http://www.invasivespeciesinfo.gov/index.shtml

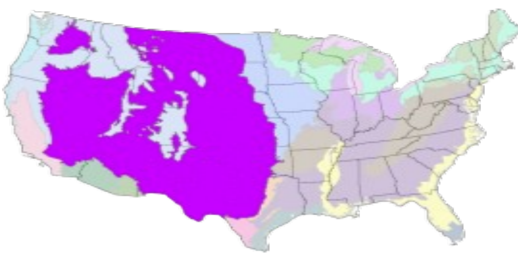
RECOVER – National Disaster/Fire Recovery

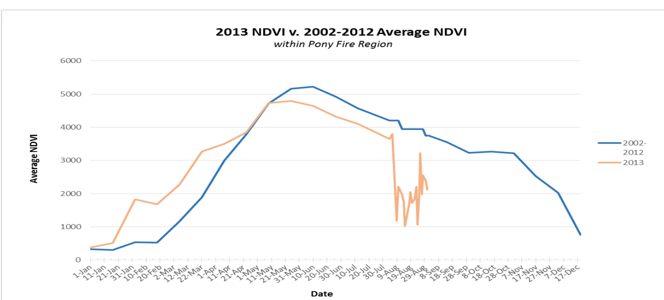
How many forest fires occurred in the US in 2014?

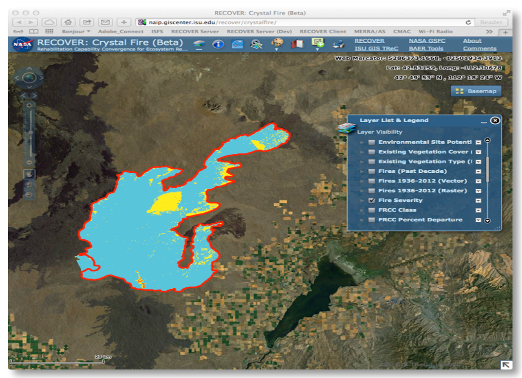
- 63,212
- http://wildland-fires.findthedata.com/

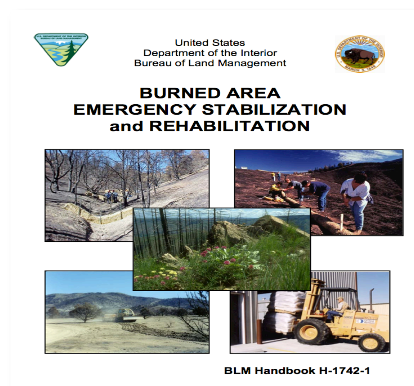
Rehabilitation Capability Convergence for Ecosystem Recovery

- John Schnase/GSFC; BLM; USGS; NPS; USDA
- Automatic aggregation of data needed for post-fire decision making
- Targets Burned Area Emergency Response (BAER) Teams
- Data Sources
  - Earth observations and derived decision products

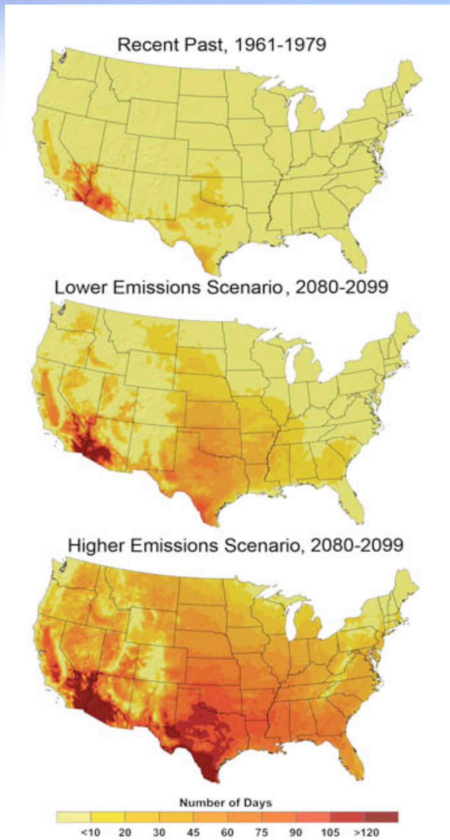




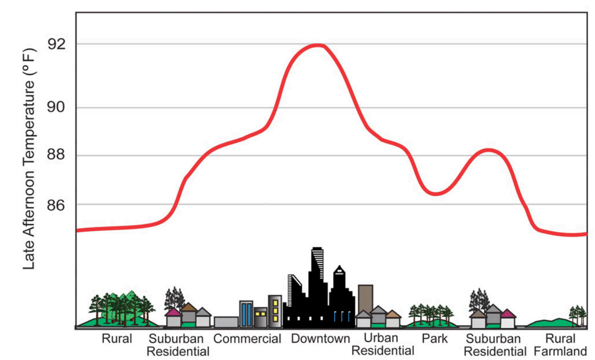




What is the impact of heat waves on health?



These charts show the number of 100-degree days per year is projected to increase.
Source: USGCRP (2009)



Heat waves in urban areas cause a large amount of health issues and loss of life for those less prepared to deal with excessive temperatures (older adults, young children, people with medical issues. How do we prepare for future heat waves?
Source: USGCRP (2009)

Studies indicate that heat waves in the Northern Hemisphere will become more likely in the coming years due to climate change:
Source: http://www.nasa.gov/topics/earth/features/warming-links.html

NASA Earth Science Division Project Won Intel Head in Clouds Challenge Award to Estimate Biomass in South Sahara

Project Goal

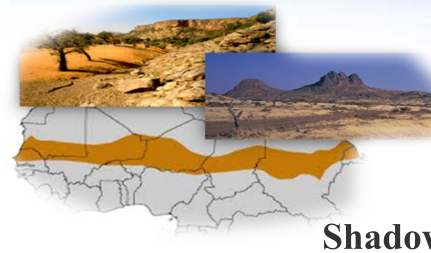
- Using National Geospatial Agency (NGA) data to estimate tree and bush biomass over the entire arid and semi-arid zone on the south side of the Sahara

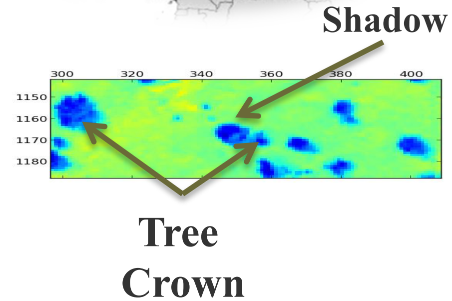
Project Summary

- Estimate carbon stored in trees and bushes in arid and semi-arid south Sahara
- Establish carbon baseline for later research on expected CO_2 uptake on the south side of the Sahara

Principal Investigators

- Dr. Compton J. Tucker, NASA Goddard Space Flight Center
- Dr. Paul Morin, University of Minnesota





NGA 40 cm imagery representing tree & shrub automated recognition

Reference: Tucker and Morin are extending earlier tree and bush mapping work published by Gonzalez, Tucker, and Sy entitled "Tree density and species decline in the African Sahel attributable to climate" in the Journal of Arid Environments in 2012.

Partners and Resources

Intel

- Professional Services and Initial Funding for AWS Resources and code optimization

Amazon Web Services (AWS)

- Compute and storage resources, support to set up the environment, consul on how to obtain the best cost solutions

Cycle Computing

- Cloud Resource Management and Data Movement Software
- Services to install and configure the software and get application running

Climate Model Data Services (CDS – GSFC Code 600)

- NGA data support

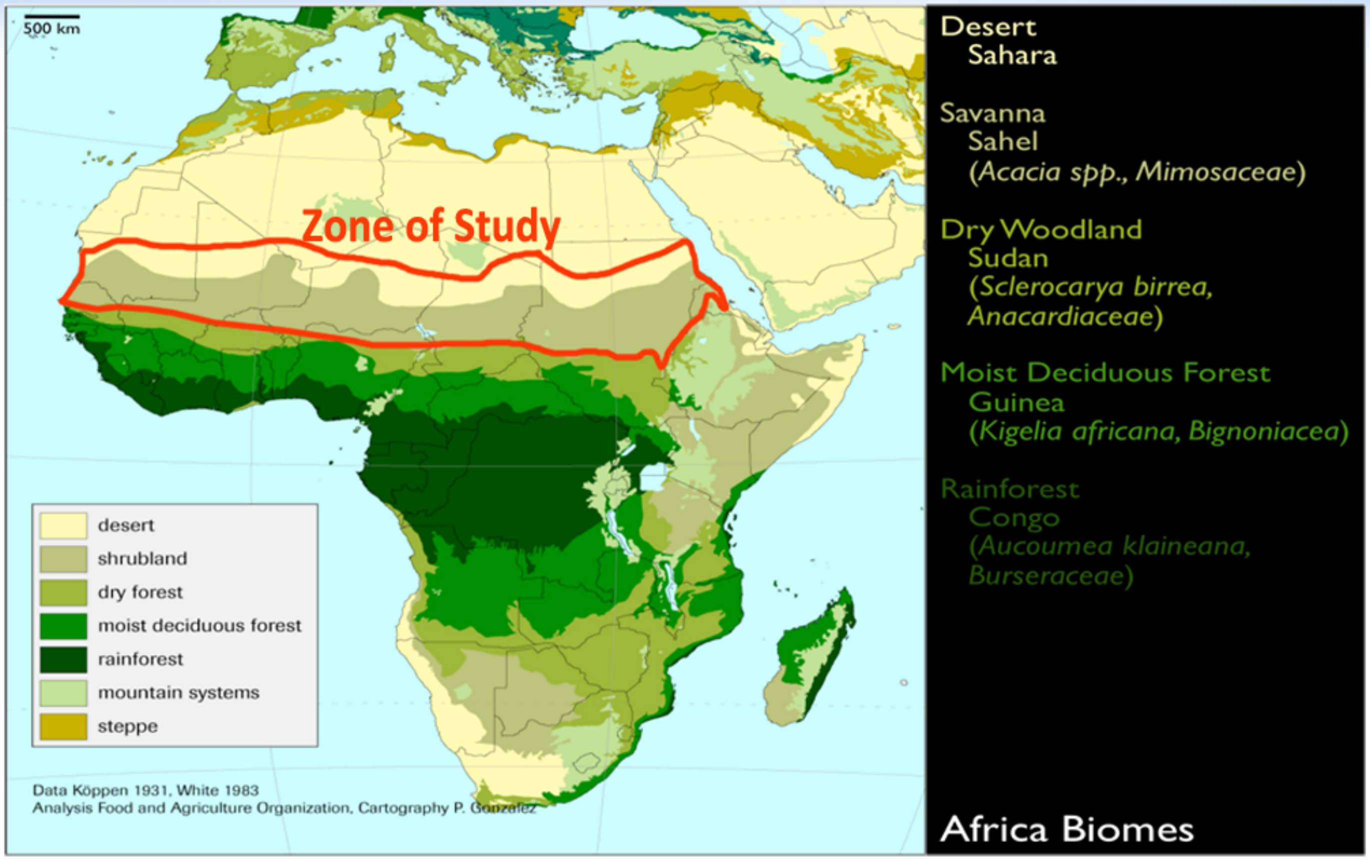
NASA Center for Climate Simulation (NCCS – GSFC Code 606.2)

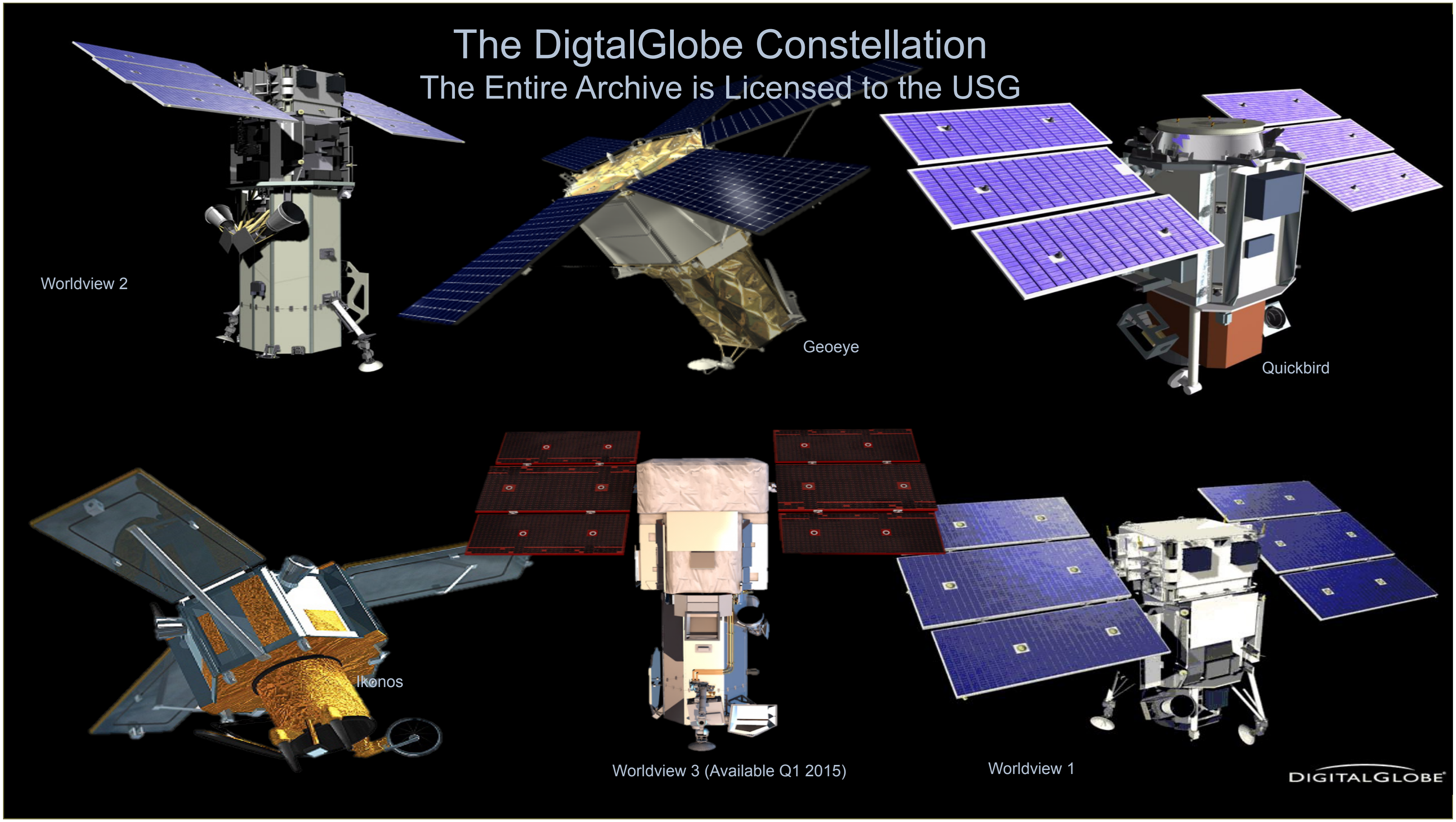
- System administration, application support, and data movement

NASA CIO

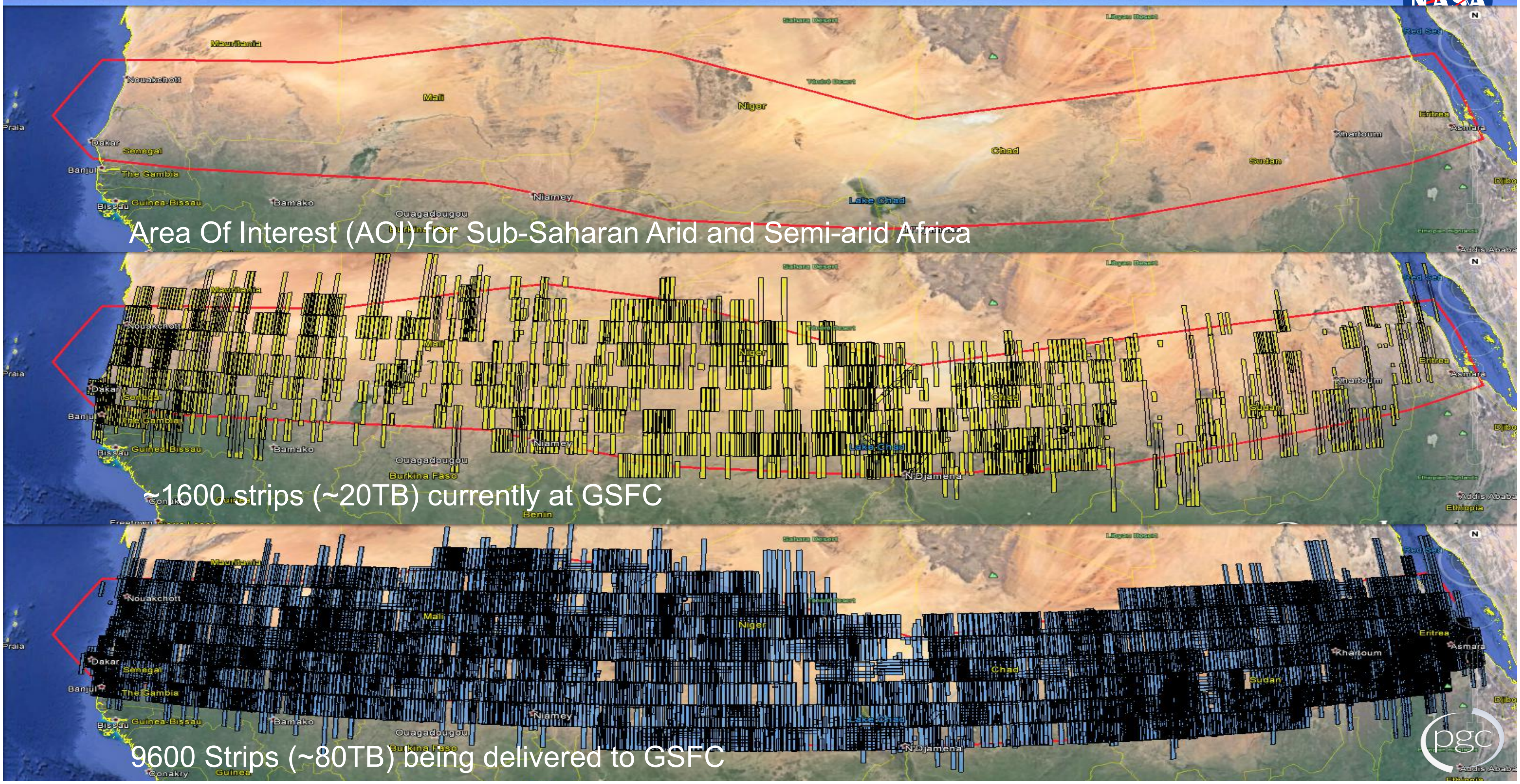
- General cloud consulting and coordination support, including networking

Desired Full Zone of Study

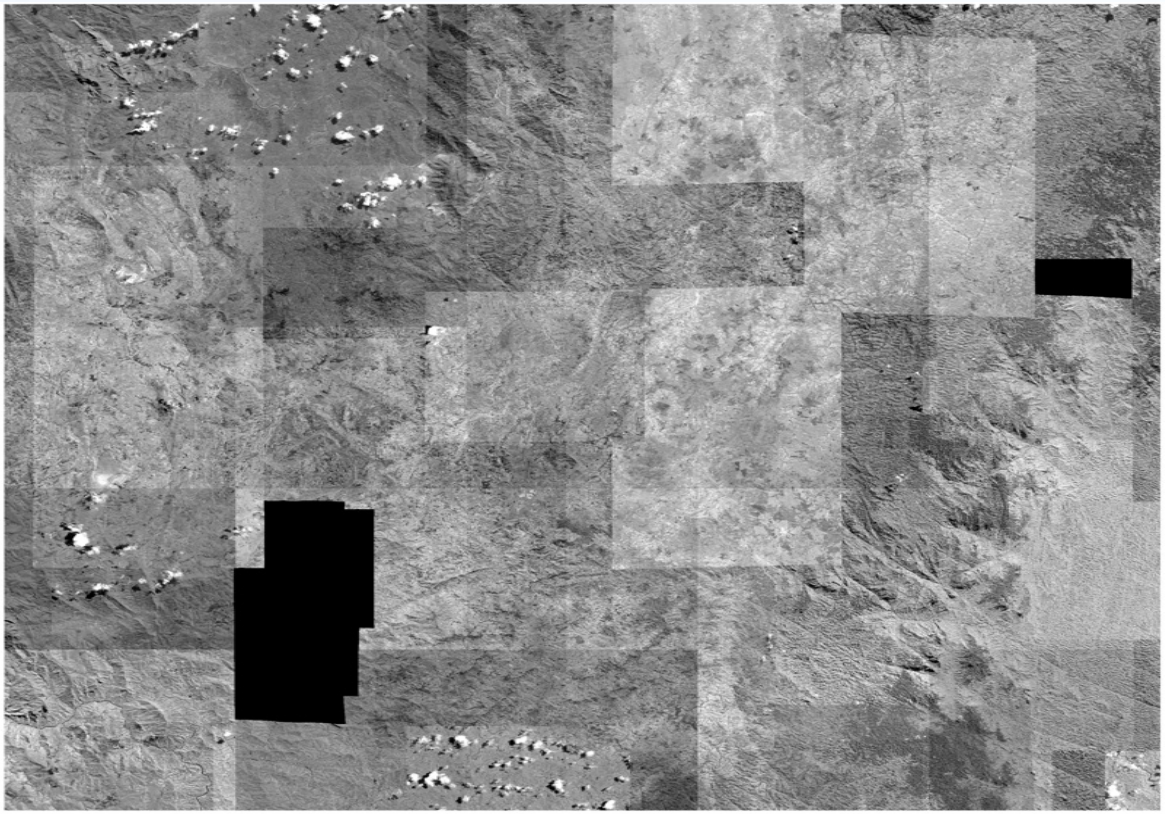



The DigtalGlobe Constellation
The Entire Archive is Licensed to the USG
Worldview 2
Geoeye
Quickbird
Ikonos
Worldview 3 (Available Q1 2015)
Worldview 1
DIGITALGLOBE

Existing Sub-Saharan Arid and Semi-arid Sub-meter Commercial Imagery

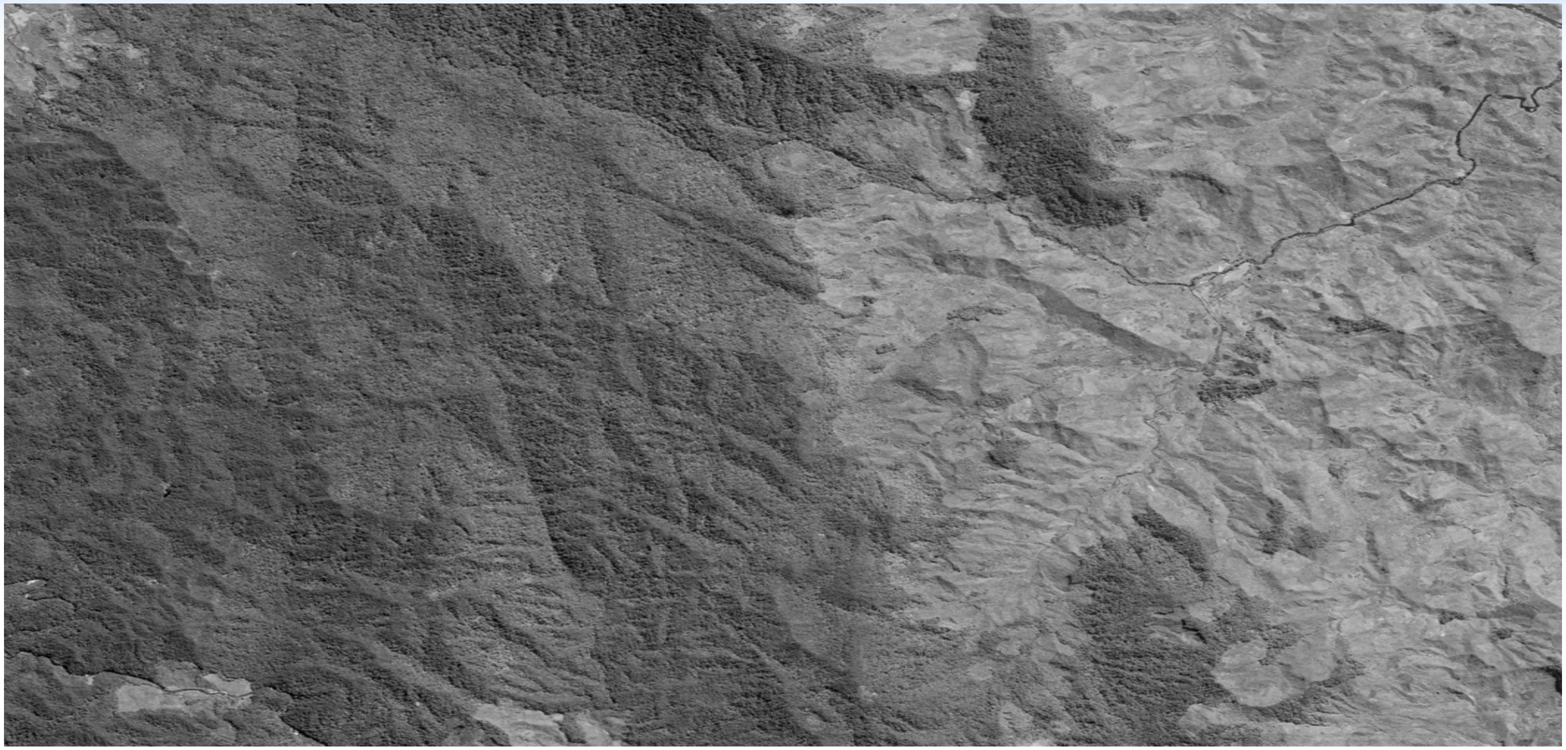


Examples of Commercial Imagery Data

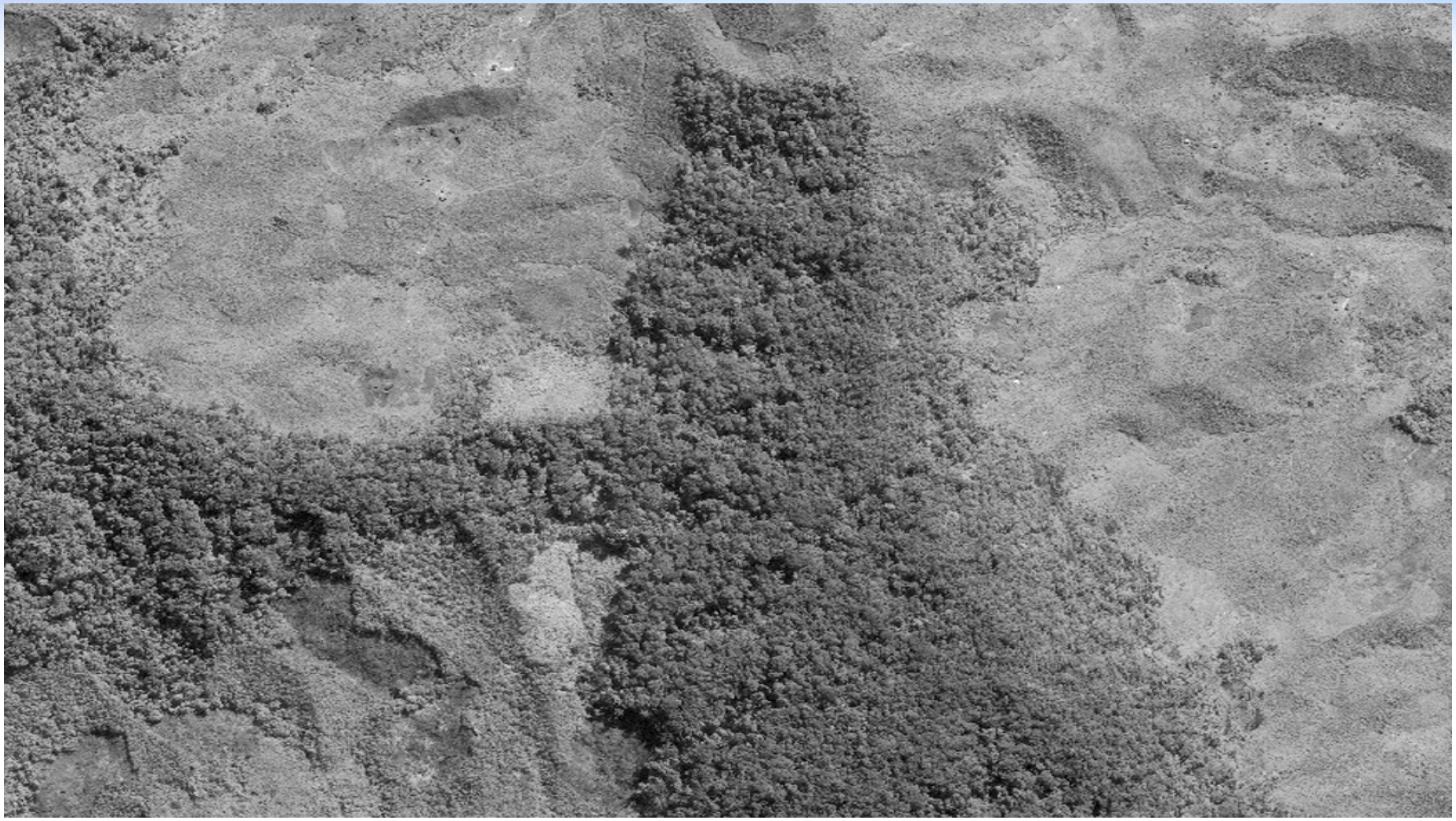


100x100 km data block of commercial satellite data

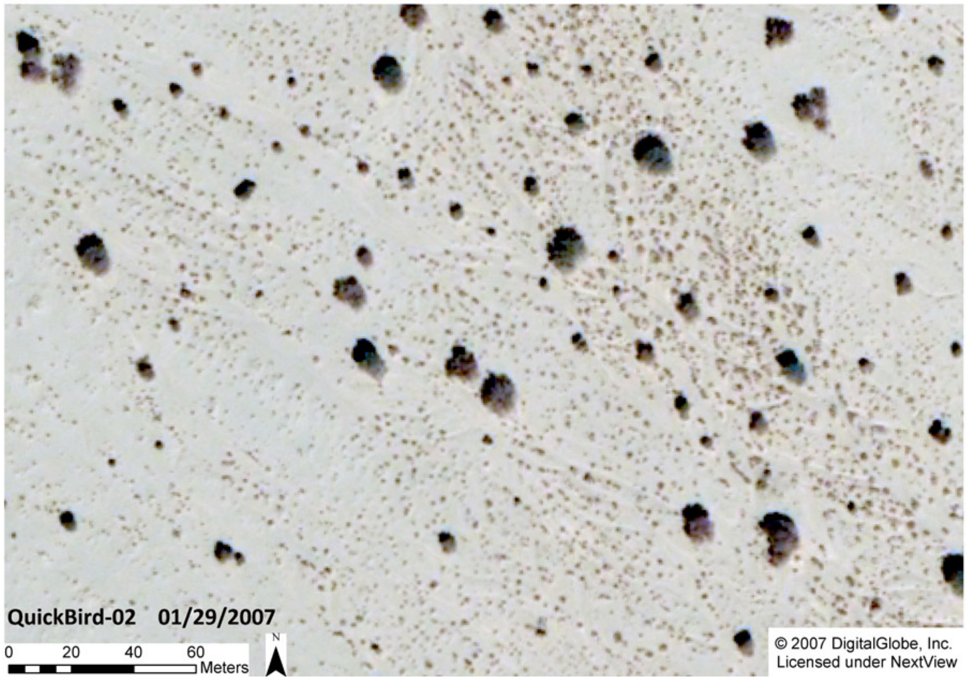
Zooming in…

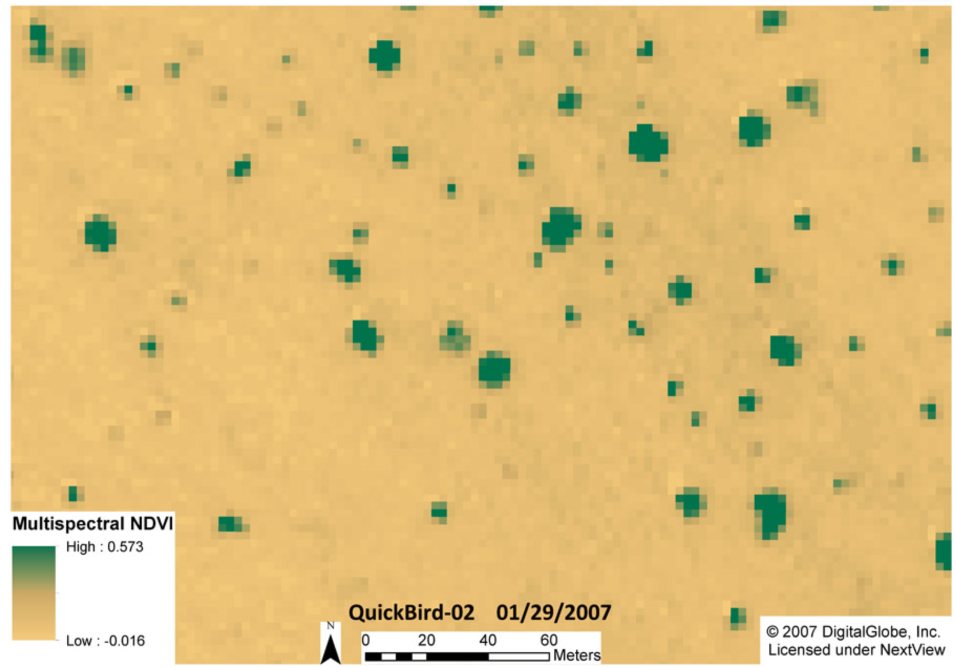


Finally, you can see trees!

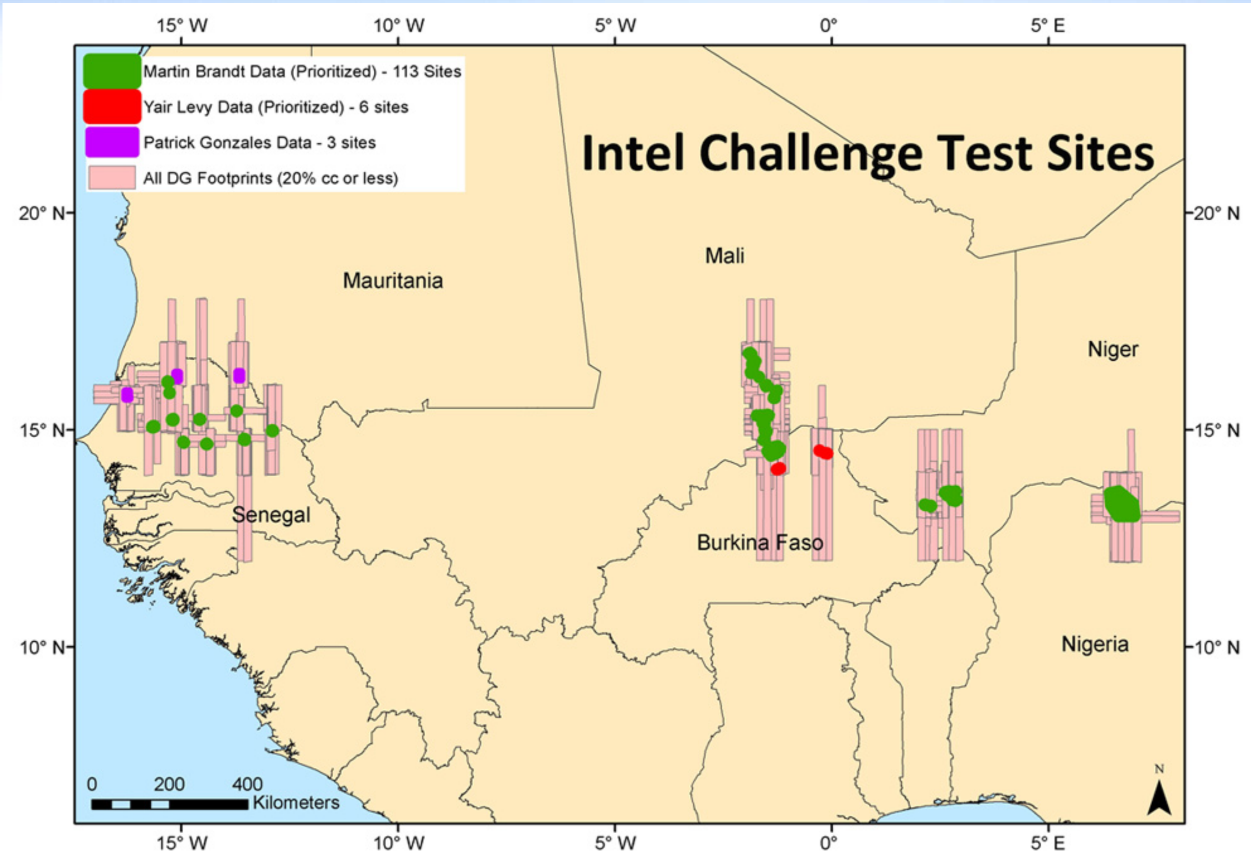


Panchromatic & Multi-spectral Mapping at the 40 & 50 cm scale





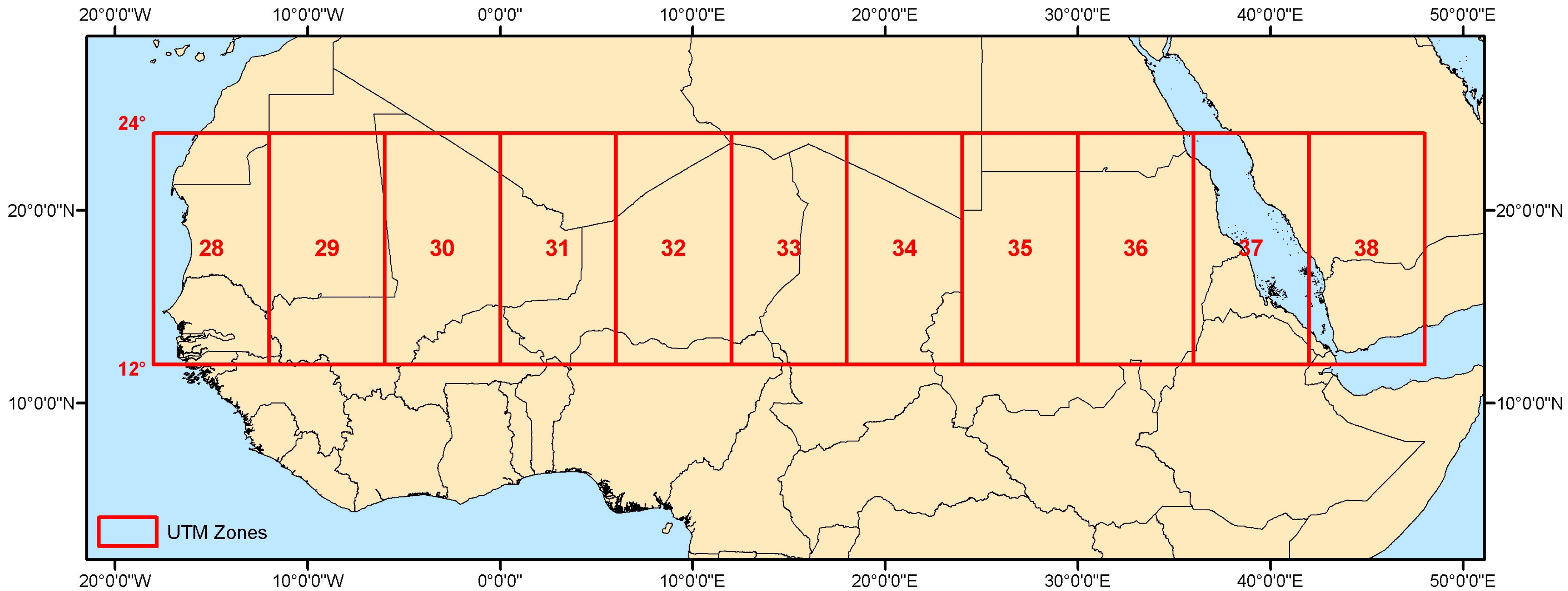
Ground Validation



Martin Brandt Measuring and Cataloguing Trees

Measuring and Cataloguing More Trees

Break the Data Down by UTM Zones



How to Break Down the Data?

Polar circumference of the Earth = 40,008 KM

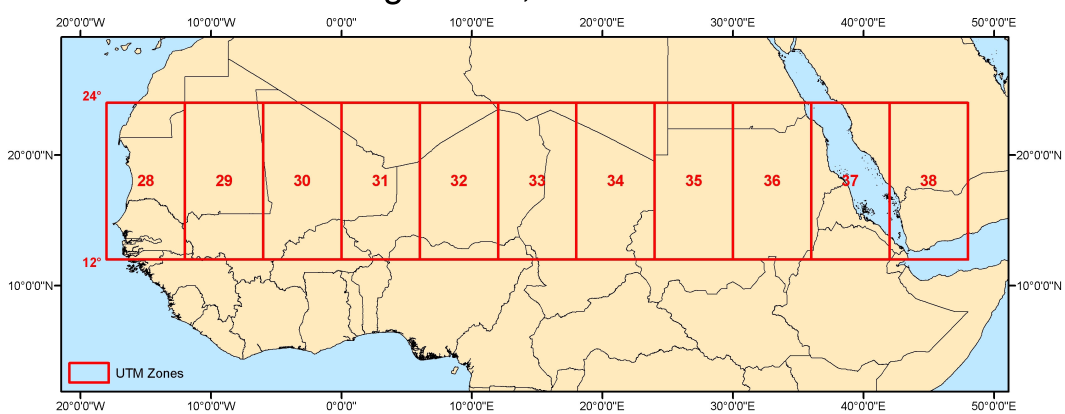
- 40,008 KM/360 latitude degrees = 111.13 KM/ latitude degree

Equatorial circumference of the Earth = 40,075 KM

- 40,075 KM/360 longitude degrees = 111.32 KM/long degree

Single UTM Zone (5.91 long degrees by 12.0 lat degrees)

- 5.91 lon degree * 111.32 KM/longitude degree = 657.9 KM
- 12 lat degree * 111.13 KM/latitude degree = 1,333.56 KM



Zones

↓

Tiles

↓

Sub-Tiles

↓

Chunks

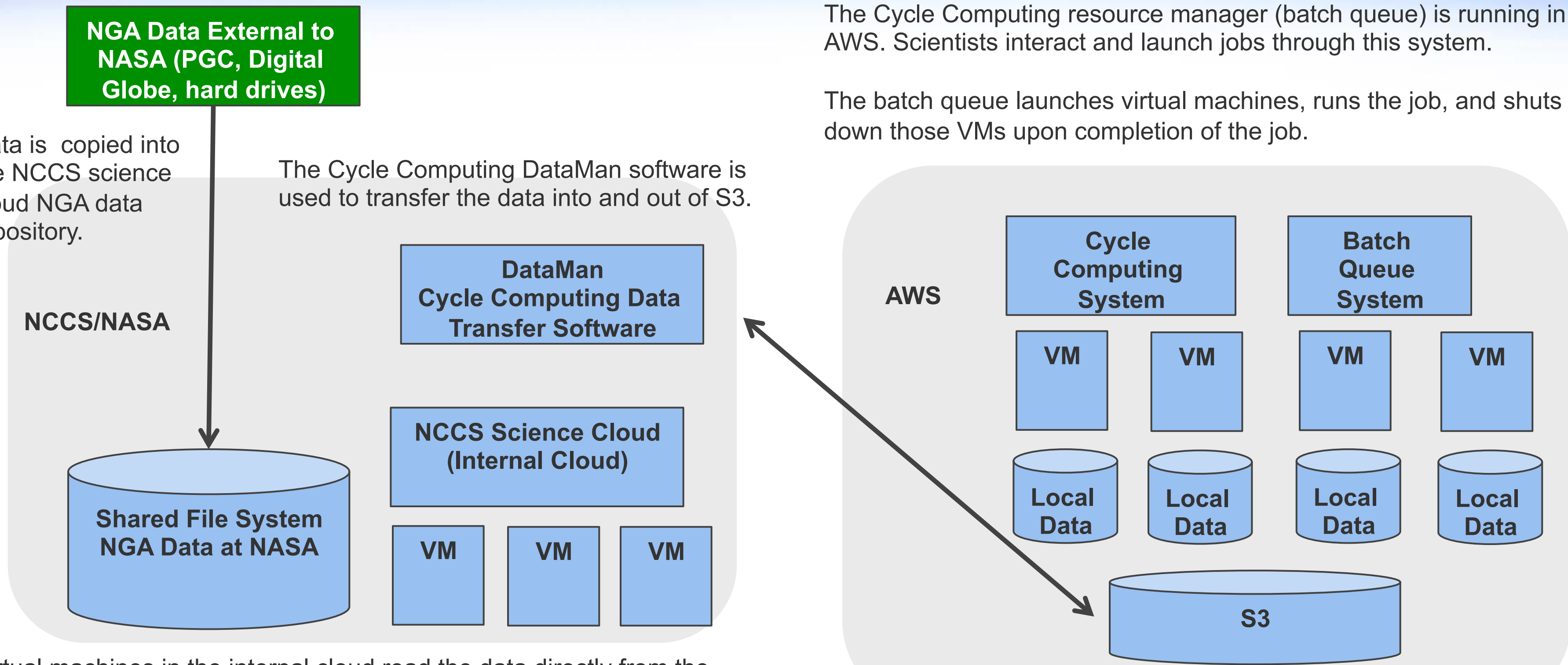
Workflow – Hybrid Cloud

The Cycle Computing resource manager (batch queue) is running in AWS. Scientists interact and launch jobs through this system.

The batch queue launches virtual machines, runs the job, and shuts down those VMs upon completion of the job.

Data is copied into the NCCS science cloud NGA data repository.

The Cycle Computing DataMan software is used to transfer the data into and out of S3.



Virtual machines in the internal cloud read the data directly from the shared disk in the NASA internal cloud. No additional data movement is required. Data is preprocessed into UTM Zones here.

UTM processed data is staged into Amazon S3. Data will be moved to the local storage of the VM's for processing. Products could be stored in S3 for transfer to the NCCS at a later time.

AWS Resources

- **AWS East**
  - Initially started using resources closer to Goddard; worried about network bandwidth
- **AWS West**
  - US West (Oregon) Region, EC2 Availability Zones: 3, Launched in 2011; Green Data Center
- **AWS Commitment to Use Renewable Energy**
  - As of April 2015, approximately 25% of the power consumed by the AWS global infrastructure came from renewable energy sources.
  - By the end of 2016, AWS plans to reach the goal of 40% power from renewable energy sources.

Network bandwidth between GSFC and AWS West is about 40-50 MB/sec.

For more information … http://aws.amazon.com/about-aws/sustainable-energy/

AWS Cluster Configuration Requirements

Category	Description	Requirement
Number of Cores	How many cores are required on a single node for the application?	1 per sub-tile
Amount of Memory (RAM)	How much memory on a node (or per core) is required for the application?	Slightly more than 4 GB per sub-tile
Operating System (O/S)	What operating system does the application need?	CentOS
Libraries/Tools/Software	What additional libraries, tools, and software are needed to be installed? Compilers? Commercial software?	None; code written in python
Parallelization	Can the application run in a parallel manner? If so, how (threaded, MPI, or multiple instances of the application)?	Inherently parallel processing of each scene and/or tile
Cluster	If the application runs in parallel across many nodes, how many nodes are required?	100’s to a few 1,000
Storage	How much storage space will be required for each run (input, intermediate, and output files)?	Total Input – 8 TB Total Output Back to NCCS – 2 TB (approx. 25% of total input)
Shared Storage	Does this storage have to be shared across all nodes?	Using S3 to move data to local VM storage; S3 used to store output

Test Runs Using AWS Spot Instances

Ran about 1/3 of UTM Zone 31 as a test with a single satellite

- 200 virtual machines using AWS spot instances
- All jobs ran successfully and were not preempted
- Each job consumed about 4.3 GB peak of memory using a single core
- All results were pushed to S3

Showed that we can scale linearly; in other words, we can compute all UTM zones in the same amount of time.

- To finish in about 6-7 hours, we would need
  - 11 UTM Zones * 4 Satellites * 600 VMs/UTM Zone = 26,400 VMs

Spot Instances

- Propose a bid price for a spot instance
- Spot instances run when your bid price exceeds the spot price
- Not guaranteed to run indefinitely
- Reduce costs by 50% to 90% from on-demand instances

Use Niger as the Test Case – UTM Zone 32

- **Input Data**
  - Currently have about 16,000 total scenes covering Niger (the data is already orthorectified)
  - Don't actually need to use all those scenes - Total input of about 8 TB or 3,120 scenes
  - Average of about 2.63 GB of data per sub-tile
- **Output Data**
  - Total output data is estimated to be 25% of the input data
  - Estimated total output is about 2 to 3 TB
  - Output data will be transferred back to the NCCS
- **Additional UTM Zones**
  - Will scale up to run all UTM zones

Okay, so how much does the compute cost?

Using AWS spot instances

- The entire test run cost $80.
- Can do an entire UTM zone for ~$250.

Cost for all 11 UTM Zones ~$2,750

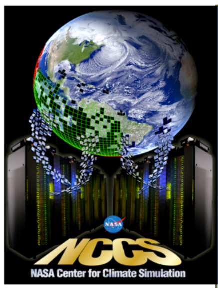
Cost for all 11 UTM Zones and all 4 satellites ~$11,000

Well, what about storage?

- There will be some costs for storage, but not much.
- Once the data is used, we will remove it from AWS.
- Only transfer out the results, which are much smaller than the inputs.

These costs are well within the budgetary constraints of typical science proposals.

Make This a Service Within the NCCS



ADAPT

High Performance Science Cloud

Big Data In

Results Out

Commercial Clouds

AWS
MS Azure
Other

Private cloud within the NCCS designed for large-scale data analytics.

Ability to burst into commercial clouds as needed depending on science requirements.

Leveraging the NASA CIO Enterprise Cloud Computing services, science projects would provide funding to a WBS in NASA to get to commercial cloud offerings.

Potentially burst into multiple commercial clouds.

Leverage the best value solution for the science application.

Currently testing MS Azure with other applications.

Vision of the Future

What is the future vision of the merging of Exascale computing with data analytics?

What can we learn from a Selective Attention Test?

Selective Attention Test References

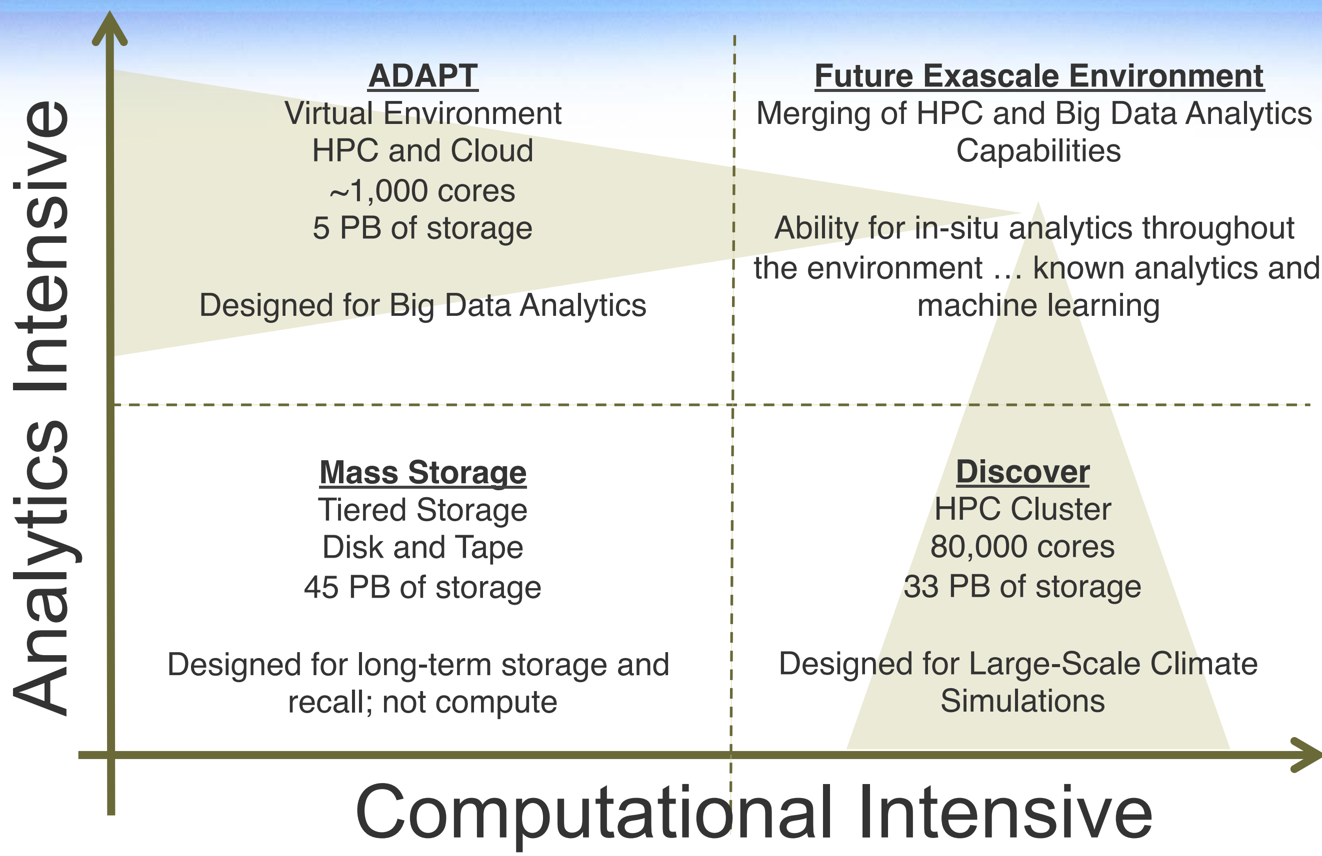
https://www.youtube.com/watch?v=vJG698U2Mvo

http://www.theinvisiblegorilla.com/

The Growth of Climate Data

35-year Reanalysis	Resolution	Data Size
MERRA and MERRA2 – Current NASA reanalysis	50 KM	400 TB
Current NASA operational resolution (working toward 13 KM resolution)	25 KM	1.6 PB
Current NOAA operational resolution; 15 of 35 years will be complete by this fall (2015 – THIS YEAR)	13 KM	6.4 PB
Cloud permitting models, still parameterized (currently have a 2 year simulation)	7 KM	26 PB
Current high resolution climate runs (currently have a 3 month simulation)	3 KM	102 PB
Resolving deep convection – currently simulate 1 model day per wall clock day (model climate in real time)	1 KM	410 PB
Cloud permitting – working toward coupled models (atmosphere, cloud, ocean, wave, ice, etc.)	0.75 KM	1.6 EB

The Future of Big Data and HPC at Exascale



https://www.whitehouse.gov/the-press-office/2015/07/29/executive-order-creating-national-strategic-computing-initiative

Thanks goes many people …

NASA

- Dr. Compton Tucker (Co-PI/GSFC)
- John David (GSFC)
- Katherine Melocik (GSFC)
- Jennifer Small (GSFC)
- Dr. Tsengdar Lee (HQ)
- Dr. Daniel Duffy (GSFC)
- Mark McInerney (GSFC)
- Hoot Thompson (GSFC)
- Garrison Vaughn (GSFC)
- Brittany Wills (GSFC)
- Scott Sinno (GSFC)
- Ray Obrien (ARC)
- Richard Schroeder (ARC)
- Milton Checchi (ARC)

University Partners

- Paul Morin (Co-PI, Univ. Minnesota)
- Claire Porter (Univ. Minnesota)
- Jamon Van Den Hoek (Oak Ridge)

Cycle Computing

- Tim Carroll
- Michael Requa
- Carl Chesal
- Bob Nordlund
- Glen Otero
- Rob Futrick

AWS

- Jamie Baker
- Jeff Layton

Special thanks to Intel for providing the initial research grant for AWS resources, and also AWS and Cycle Computing for their continued support. There are many others who have contributed… My apologies for those I missed. Another special thanks for many of the above that have to suffer through my conference calls every other week!